Journal of Open Humanities Data

Converting the British Library's Catalogue of British and Irish Newspapers into a Public Domain Dataset: Processes and Applications

DATA PAPER

YANN RYAN
LUKE MCKERNAN
*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

This paper describes the production of a title-level list of British, Irish, British Overseas Territories and Crown Dependencies newspapers (1621–2019) held by the British Library, and its potential for reuse and research. The data was extracted from the British Library's catalogue of over 24,000 British and Irish newspaper titles, cleaned, and published on the British Library Research Repository, an open access repository for the research produced by staff and research associates of the British Library. Bespoke versions of the data have been made available to specialist users, notably the British Library/Alan Turing Institute's 'Living with Machines' project, enabling greater historical analysis of nineteenth-century British news and selective digitisation.

CORRESPONDING AUTHOR:
Yann Ryan
Dept. of English, Queen Mary, University of London, London, UK
y.c.y.ryan@qmul.ac.uk

KEYWORDS:
newspapers; digital humanities; metadata; newspaper history; British Library; archives

TO CITE THIS ARTICLE:
Ryan, Y., & McKernan, L. (2021). Converting the British Library's Catalogue of British and Irish Newspapers into a Public Domain Dataset: Processes and Applications. *Journal of Open Humanities Data*, 7: 1, pp. 1–5. DOI: *https://doi.org/10.5334/johd.23*

OVERVIEW

REPOSITORY LOCATION

DOI: *https://doi.org/10.23636/1136*

CONTEXT

Produced as part of the British Library's Heritage Made Digital newspaper project, digitising historical newspapers and exploring options for creative re-use of newspaper data (*https://blogs.bl.uk/thenewsroom/2019/01/heritage-made-digital-the-newspapers.html*).

METHOD

STEPS

The original data comes from the British Library's catalogue of world newspapers (there is no separate newspaper catalogue, but all newspaper titles are included in the British Library's Aleph management system and discoverable via its integrated catalogue at *https://explore.bl.uk*). This has been built up over c. 150 years of collecting of newspapers, a collection that now comprises some 35,000 titles or 60 million issues dating from 1619 to the present day. The collection of British and Irish titles (around two-thirds of the entire newspaper collection) runs from 1621 to the present day. It is not absolutely complete, but most titles published from the 1840s onwards are held, and effectively all titles are held from 1869 onwards, when Legal Deposit was introduced, by which publishers are required to send one copy of each newspaper issue to the British Library. There are a few omissions (either entire titles or gaps in the run of a title), while for reasons of space usually only one edition of an issue has been taken by the Library since 1869.

The newspaper catalogue is at title-level, with changes in a newspaper title and regional variants resulting in a new catalogue record, and often a new catalogue record where there has been a change in format (i.e., microfilm or digital copies). Over the long period of collecting, inevitable inconsistencies and gaps in the metadata have built up. The titles in the dataset are exactly as reflected in the catalogue, following the British Library cataloguing practice of the times when the titles were acquired.

The data was extracted from the British Library catalogue (through the Aleph Integrated Library System) by the Collection Metadata team. Aleph stores metadata relating to the newspaper collection, in the form of MARC records, in a number of fields, often with complicated holding information. The Collections Metadata team at the British Library extracted years of publication from the free-text information about date holdings, which is published alongside the original holdings field. This was then aligned to data from a separate Master Negative Database of microfilm copies of newspaper print originals (newspaper titles linked by system ID numbers), as well as an up-to-date list of digital holdings on the British Newspaper Archive, which hosts digitised newspapers from the British Library collection (*https://www.britishnewspaperarchive.co.uk*).

The initial data extraction was followed by a process of cleaning the extracted data by the News Collections team. We manually adjusted some 2,000 holdings records, mostly where years had not been extracted or there were inconsistencies surrounding place of publication—either where alternative spellings or punctuation had been used, or to ensure that the entire dataset used the same set of UK county boundaries. For this, and for joining the initial dataset to the microfilm and digital holdings, custom scripts for cleaning and extracting the data were developed using R, before exporting to the final .csv format.

SAMPLING STRATEGY

To produce a reusable dataset, the decision was made to limit this to British and Irish newspapers, where there were fewer complications with the data, such as dealing with languages other than English, the need for research into the history of some titles, or requiring consultation with other British Library curators in relevant area studies. A complete listing of all titles in the newspaper collection will be a follow-up project, scheduled to take place in 2021.

DOI: 10.5334/johd.23

QUALITY CONTROL

Not Applicable.

DATASET DESCRIPTION

OBJECT NAME

British and Irish Newspapers: A title-level list of British, Irish, British Overseas Territories and Crown Dependencies newspapers held by the British Library

FORMAT NAMES AND VERSIONS

Excel; CSV; plaintext

CREATION DATES

2016-01-22–2019-11-18

DATASET CREATORS

Danskin, Alan – Collection Metadata Standards Manager (British Library)

Lester, Stephen – Curator, Newspaper Collections (British Library)

McKernan, Luke – Lead Curator, News and Moving Image (British Library)

Ryan, Yann – Curator, Newspaper Data (now post-doctoral researcher, Queen Mary University of London)

LANGUAGE

English

LICENSE

CC0

REPOSITORY NAME

British Library Research Repository

PUBLICATION DATE

2019-11-18

REUSE POTENTIAL

The past six years or so have seen the rise of 'Collections as Data': the idea that metadata from holdings of cultural heritage collections can function as data to be analysed in its own right (see Collections as Data National Forum, 2018, for a definition and discussion of the term). Tim Sherratt, for example, has used the metadata from the National Library of Australia's Trove digitised newspaper collection to undertake historical analyses (Sherratt, 2019).

As newspapers are digitised, detailed metadata are produced in tandem, providing issue-level details on the place and date of publication, which can then be exploited by researchers (Fyfe, 2016). However, this only relates to the portion of the collection which has been digitised, currently consisting of just over 8% of the entire British Library newspaper collection of 450 million pages. Up until now, no easily available survey of the *print* holdings of the Library has been available to researchers. We see the main reuse potential of this dataset as four-fold:

Firstly, the list can be used in conjunction with the physical holdings and the Library's Explore catalogue (*https://explore.bl.uk*) as a general finding aid, or to narrow down one's search to a specific corpus of newspapers. While Explore is already an excellent search tool, this list aids discovery by enabling easy browsing by date and location.

Secondly, it opens up newspaper data to the non-specialist. We purposely standardised and simplified the data fields so that users could take advantage of the filtering, sorting and graphing functions in software such as Excel or Google sheets.

Thirdly, it allows for geographical and diachronic analyses of the British and Irish newspaper industries, allowing for easy production, for example, of time-series statistics on the establishment of new titles, or of maps of individual 'hotspots' of newspaper growth on a county or city level. While geographic coordinates were beyond the scope of the dataset, a code to accurately georeference the structured data is being developed specifically for use with the title list (Ryan et al., 2020).

Finally, understanding the print collection helps us to understand the digitised portion in context. Researchers across the world now use the data from the British Library's digitised newspaper collection for historical research. Many of these projects employ large-scale text mining or image analytics over the entire collection to make broad historical claims: previous projects, for instance, have used the corpus to estimate dates when electricity took over from horses, or to analyse 'subjective well-being' (Lansdall-Welfare et al., 2017; Hills, Proto, Sgroi, & Seresinhe, 2019). However, it is also recognised that these types of claims must be understood in terms of the idiosyncrasies existing in the digitised collection. The corpus ultimately only represents a fraction of the entirety of the Library's newspaper collection and has not been produced to be particularly systematic or representative (Shaw, 2005). This list helps to contextualise the data in the digitised collection. A project undertaking text mining, for example, may adjust the weighting methods if one understands the proportions of the print holdings of a particular place that each digitised newspaper represents. The 'Living with Machines' project — a British Library and Alan Turing Institute initiative using data analysis to understand the lived experience of the nineteenth century — is already using a version of this list to carry out a 'topographical survey' of the digitised newspaper collection (Vane, 2020).

ACKNOWLEDGEMENTS

The authors would like to thank the British Library News Collections team, British Library Digital Scholarship, and the researchers of the 'Living with Machines' project for their help with and feedback on this dataset.

FUNDING INFORMATION

British Library grant-in-aid.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Ryan: Conceptualization; Investigation; Data curation; Writing – original draft

McKernan: Conceptualization; Investigation; Supervision; Resources; Project Administration; Writing – review & editing

AUTHOR AFFILIATIONS

Yann Ryan *orcid.org/0000-0003-1878-4838*
Department of English, Queen Mary, University of London, London, UK

Luke McKernan *orcid.org/0000-0002-6285-295X*
Newspaper Collections, British Library, London, UK

REFERENCES

Collections as Data National Forum. (2017). *The Santa Barbara Statement on Collections as Data*. Retrieved November 20, 2020, from *https://collectionsasdata.github.io/statement/*

Fyfe, P. (2016). An archaeology of Victorian newspapers. *Victorian Periodicals Review*, *49*(4), 546–577. DOI: *https://doi.org/10.1353/vpr.2016.0039*

Hills, T. T., Proto, E., Sgroi, D., & **Seresinhe, C. I.** (2019). Historical analysis of national subjective wellbeing using millions of digitized books. *Nature Human Behaviour*, *3*(12), 1271–1275. DOI: *https://doi.org/10.1038/s41562-019-0750-z*

Lansdall-Welfare, T., Sudhahar, S., Thompson, J., Lewis, J., FindMyPast Newspaper Team, & **Cristianini, N.** (2017). Content analysis of 150 years of British periodicals. *Proceedings of the National Academy of Sciences*, *114*(4), E457–E465. DOI: *https://doi.org/10.1073/pnas.1606380114*

Ryan, Y., Ardanuy, M. C., Van Strien, D., Hosseini, K., Beelen, K., Hetherington, J., McDonough, K., McGillivray, B., Ridge, M., Vane, O., & **Wilson, D.** (2020). Using Smart Annotations to Map the Geography of Newspapers. ***Paper presented at DH2020***, Ottawa, Canada (held online). *https://dh2020.adho.org/wp-content/uploads/2020/07/532_Usingsmartannotationstomapthegeographyofnewspapers.html*

Shaw, J. (2005). 10 Billion Words: The British Library Newspapers 1800–1900 Project: Some Guidelines for Large-Scale Newspaper Digitisation. ***Paper presented at IFLA***, Oslo. *https://origin-archive.ifla.org/IV/ifla71/papers/154e-Shaw.pdf*

Sherratt, T. (2019). *From Collection Search to Collections as Data*. Retrieved from *http://doi.org/10.5281/zenodo.3551405*

Vane, O. (2020). *Press Picker: Visualising Formats and Title Name Changes in the British Library's newspaper holdings*. Retrieved from *https://livingwithmachines.ac.uk/press-picker-visualising-formats-and-title-name-changes-in-the-british-librarys-newspaper-holdings/*

TO CITE THIS ARTICLE:
Ryan, Y., & McKernan, L. (2021). Converting the British Library's Catalogue of British and Irish Newspapers into a Public Domain Dataset: Processes and Applications. *Journal of Open Humanities Data*, 7: 1, pp. 1–5. DOI: *https://doi.org/10.5334/johd.23*

Published: 22 January 2021

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.